IBM Data Science Capstone Project

Introduction

This capstone project is part of the requirements to successfully completing the IBM Data Science Professional Certificate. In this project, I am going to focus on Toronto, Canada, where I am going to use all what I have learned throughout this course to suggest suitable locations for a Japanese restaurant based on geographical locations only. Assume that the demand for Japanese cuisines is rising in Toronto, and it becomes a battle of location for prospective Japanese restaurants. I am going to leverage on the Foursquare API to get location data in order to suggest possible gold mine locations for a client who wants to open a Japanese restaurant.

Business Problem

The aim of this project is to suggest suitable locations for a Japanese restaurant in Toronto, Canada. A lot of data science tools and techniques are going to be used along with different machine learning techniques such as clustering. All these are put in place in order to answer the business problem:

"Where would a potential Japanese restaurant owner consider setting up shop in Toronto, Canada?"

Target Audience

In this project, the target audience is an interested potential restaurant owner who wants to open a Japanese restaurant in Toronto, Canada. From the assumption made above, this exercise will predict a good location where it would be easiest to establish a good customer base and patronisers for the restaurant. Think of it as an early bird scenario. This will definitely be of interest to the stakeholder.

Data to be used

In this project, we will need the following data:

1. **A list of neighbourhoods and their postal codes in Toronto, Canada**. This dataset contains three columns which include Postal code, Boroughs and Neighbourhoods. The data than can be extracted from this dataset are rows with a defined Borough.

2. **Geographic coordinates of these neighbourhoods (longitudes and latitudes).** This dataset contains a list of the longitudes and latitudes of the respective postal codes in Canada. We can extract the longitudes and latitudes of the postal codes

related to Toronto so we can be able to work with them when using Folium and Foursquare API

3. **Venue data of Japanese related restaurants in Toronto,** so we can determine the regions with fewer number of Japanese restaurants. This data will contain a list of all the places with in a given radius, so we can use this to get our information of the shops and stores around. The major data we are looking for is the venue category of each shop

Acquiring the data

The list of neighbourhoods will be scrapped from the Wikipedia page of Neighbourhoods in Canada.

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M,

The geographical coordinates of each neighbourhood will be downloaded from the following link,

http://geoawesomeness.com/developers-up-in-arms-over-google-maps-api-insane-price-hike/,

I have also ensured that this data is authentic.

The Foursquare API will be used to get information on the Japanese related restaurants in the state of Toronto.

Methodology

The first thing I did was to import Python libraries such as Pandas and Numpy, for handling the data, BeautifulSoup and requests to be able to get data off the web, and finally folium for displaying geospatial data. After this, I scrapped the Wikipedia for the postal code data using BeautifulSoup and requests to get the raw data from Wikipedia. The raw data consisted of three columns; the postal code, the borough and the neighbourhood.

After this I used the Pandas package to clean the data by removing unassigned boroughs. The next thing I did after this was to import the geographical coordinates of each of these postal codes to a pandas frame, after which I joined both data frames into one consisting of five columns; postal code, borough, neighbourhood, latitude and longitude. I then used folium to generate a geospatial representation of the neighbourhoods.

Next thing I did was to use the Foursquare API to get information about the top 200 venues within a 500 metre radius, after which I carried out exploratory data analysis on the dataset of venues that was generated. The type of data I was able to get from the Foursquare API were establishment name, category, longitude and latitude of the venue. With this data, I was able to check the number of unique locations in the area. I then analysed the data by determining the frequency of occurrence of each venue category. All this was done in preparation for clustering.

After this I narrowed my search down to just Japanese restaurants in the area. I then carried out the unsupervised machine learning algorithm clustering using k-means clustering on the project, using a value of three clusters. I used k-means clustering because it has the ability to form unusually shaped clusters, having high tolerance for noise. This makes it suitable for this project.

I was able to cluster the neighbourhoods of Toronto, based on the occurrence of Japanese restaurants in the area. Based on the results, I was able to draw inferences.

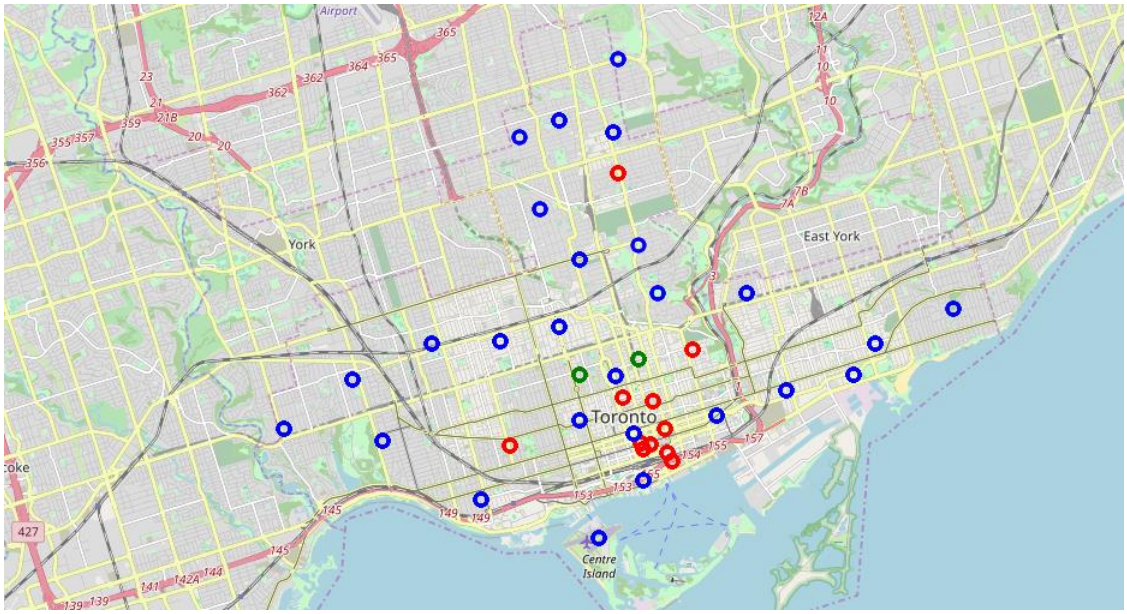
Results

Out[34]:

	Postal code	Borough	Neighborhood	Latitude	Longitude
0	M4E	East Toronto	The Beaches	43.676357	-79.293031
1	M4K	East Toronto	The Danforth West / Riverdale	43.679557	-79.352188
2	M4L	East Toronto	India Bazaar / The Beaches West	43.668999	-79.315572
3	M4M	East Toronto	Studio District	43.659526	-79.340923
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790
34	M6P	West Toronto	High Park / The Junction South	43.661608	-79.464763
35	M6R	West Toronto	Parkdale / Roncesvalles	43.648960	-79.456325
36	M6S	West Toronto	Runnymede / Swansea	43.651571	-79.484450
37	M7A	Downtown Toronto	Queen's Park / Ontario Provincial Government	43.662301	-79.389494
38	M7Y	East Toronto	Business reply mail Processing CentrE	43.662744	-79.321558

The table above shows the result when the cleaned Wikipedia and geospatial coordinates are combined to one data frame

The diagram above shows the distribution of neighbourhoods in Toronto, Canada.



The results from the k-means algorithm show that we can classify the Toronto neighbourhoods into three, based on how many Japanese restaurants are in each neighbourhood. Cluster 1 is in red, cluster 2 is in blue and cluster 3 is in red

Discussions

From the results of the code, Cluster 0 is around the area Underground city and Garden district. Cluster 1 is around Richmond and Grange Park. Cluster 2 is around Church and Wellesley. It is obvious that cluster 0 has the highest number of Japanese restaurants while cluster 1 has the least number of restaurants. In general, each cluster has at least three Japanese restaurants.

Recommendations

Cluster 1 looks like an ideal place to open a Japanese restaurant, because of its close proximity to commercial areas, and more importantly because of the small number of already existing Japanese Restaurants. Cluster 2 would have been a suitable location, but its location is not as profitable as that of cluster 1.

Conclusion

Cluster 1 is a more profitable location for the stakeholders to open a Japanese Restaurant.